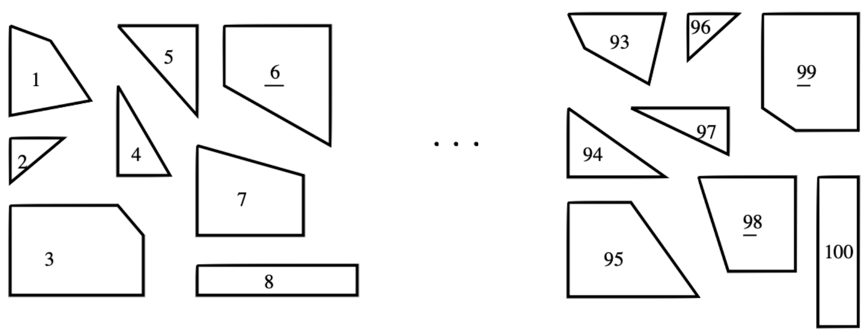
Random sampling plans

```r
## Set this up for your own directory
imageDirectory <- "MyAssignmentDirectory/img"  # e.g. in current "./img"
dataDirectory <- "MyAssignmentDirectory/data"  # e.g. in current "./data"
path_concat <- function(path1, path2, sep="/") paste(path1, path2, sep = sep)
```

---

47 marks

Consider the study population $\mathcal{P}_{Study}$ of $N = 100$ blocks of uniform thickness and density (all blocks were cut from the same opaque plastic sheet of about 5 mm thickness), but have different shapes such as those shown below:



Data on this population of 100 blocks are available as an R data set `blocks`. This data set has four variates: block `id` number, `weight` in grams, `perimeter` in centimetres, and the `group` of the block (being either `A` or `B`). It can be loaded from the assignment data directory as follows:

```r
load(path_concat(dataDirectory, "blocks.rda"))
head(blocks, n = 3)
```

```
##   id weight perimeter group
## 1  1     55        32     B
## 2  2     35        27     B
## 3  3     35        25     A
```

In this question, you will investigate different sampling plans and estimation procedures.

a. Simple random sampling.

   i. *(4 marks)* Collect the sample average block weight from each of 1000 samples, where each sample consists of 10 blocks selected at random (without replacement) from all 100 blocks.

      Before sampling, `set.seed(314159)`

      Save the results on the R variable `randomSampleAves`.

      Show your code.

   ii. *(3 marks)* Using `randomSampleAves`, estimate the sampling bias, the sampling variability, and the sampling mean squared error of this sampling plan.

      Show your code.

   iii. *(3 marks)* Construct a (suitably labelled) histogram of the sample **errors** from this sampling plan.

      Use `xlim = c(-20,20)`.

      Add a vertical red dashed line of `lwd = 2` at the average error.

      Show your code.

b. Stratified random sampling.

   i. *(4 marks)* Collect the sample average block weight from each of 1000 samples, where now each sample consists of 5 blocks selected at random (without replacement) from each of group "A" and group "B".

      Before sampling, `set.seed(314159)`

      Save the results on the R variable `stratifiedSampleAves`.

      Show your code.

   ii. *(3 marks)* Using `stratifiedSampleAves`, estimate the sampling bias, the sampling variability, and the sampling mean squared error of this sampling plan.

      Show your code.

   iii. *(3 marks)* Construct a (suitably labelled) histogram of the sample **errors** from this sampling plan.

      Use `xlim = c(-20,20)`.

      Add a vertical red dashed line of `lwd = 2` at the average error.

      Show your code.

c. Regression estimators. In this question, we suppose that we know something about the population of blocks. In particular, suppose we know that the average perimeter of all 100 blocks is `mean(blocks$perimeter)` = 26.27.

   We also understand that there is some relationship between `perimeter` and `weight` in this population.

   i. *(4 marks)* Here 1,000 samples of 10 blocks are to be selected at random (without replacement) from all 100 blocks. For each sample of 10 blocks, construct a straight line fit of the `weight` on `perimeter`. Then use this fit to predict the mean weight of the population when the `perimeter` is the actual average perimeter of all 100 blocks. Collect all 1,000 regression estimates.

      Before sampling, `set.seed(314159)`

      Save the results on the R variable `regressionEstimates`.

      Show your code.

ii. *(3 marks)* Using `regressionEstimates`, estimate the sampling bias, the sampling variability, and the sampling mean squared error of this sampling plan.

    Show your code.

iii. *(3 marks)* Construct a (suitably labelled) histogram of the sample **errors** from this sampling plan.

    Use `xlim = c(-20,20)`.

    Add a vertical red dashed line of `lwd = 2` at the average error.

    Show your code.

iv. *(2 marks)* Is the straight line model used in this question "true"? Is it useful? Explain your answers.

d. A number of graduate data science students were asked to view the entire collection of 100 blocks and to choose 10 whose average weight they believed came close to matching that of all 100. The sample units selected are recorded in another file, `judgmentSamples.csv`. These can be loaded from the assignment data directory as follows:

```
students <- read.csv(path_concat(dataDirectory, "judgmentSamples.csv"))
head(students, n = 3)
```

```
##   studentID first second third fourth fifth sixth seventh eighth ninth tenth
## 1      5086    12     18    17     11    15    20      14     13    16    18
## 2      3848    34     35    70     56    32    14       5     88    81    73
## 3      6656    14     34    41     29    32    55      74     40    16    70
```

There were a total of 33 students and hence 33 samples selected.

In this question, we compare the **judgment** sampling plan of the students with that of the random sampling plans considered above when only 33 samples of size 10 are selected.

i. *(2 marks)* Gather together the average block weights of the student judgment samples. Save the results on the R variable `judgmentAves`.

    Print the average of these averages.

    Show your code.

ii. *(8 marks)* Using `judgmentAves` and only the first 33 entries of each of `randomSampleAves`, `stratifiedSampleAves`, and `regressionEstimates`, construct four histograms one above the other (in the same display, use an appropriate `par()`) one for each of these sets of results.

    Make sure each histogram is labelled appropriately.

    Use the same `xlim = c(20, 50)`, `ylim = c(0, 15)`, and `breaks = seq(20, 50, 2)` for each histogram.

    On each histogram add a vertical red dashed line (with `lwd = 2`) at the true population average weight of all 100 blocks.

    On each histogram add a vertical "steelblue" **solid** line (with `lwd = 2`) at the average of all 33 sample estimates.

    On each histogram, add a legend indicating which vertical line is which.

    Show your code.

e. *(5 marks)* Comment on the relative merits of the four sampling plans. Which would you most recomment? Which least?